08421 Abstracts Collection
Uncertainty Management in Information Systems
— Dagstuhl Seminar —

Ch. Koch[1], Birgitta König-Ries[2], Volker Markl[3] and Maurice van Keulen[4]

[1] Cornell University, USA
[2] Universität Jena, Germany
[3] IBM Almaden Center - San José, USA
[4] University of Twente, The Netherlands

Abstract. From October 12 to 17, 2008 the Dagstuhl Seminar 08421 "Uncertainty Management in Information Systems " was held in Schloss Dagstuhl – Leibniz Center for Informatics. The abstracts of the working group reports, the plenary and session talks given during the seminar as well as those of the shown demos are put together in this paper.

Keywords. Uncertainty management

08421 Executive Summary – Uncertainty Management in Information Systems

This executive summary provides a brief overview of the topic, the organization, and the outcome of the Dagstuhl Seminar on Uncertainty Management in Information Systems.

Joint work of: Koch, Christoph; König-Ries, Birgitta; Markl, Volker; van Keulen, Maurice

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2009/1940

1 Working Group Reports

08421 Working Group: Classification, Representation and Modeling

This report briefly summarizes the discussions carried out in the working group on classifffication, representation and modeling of uncertain data. The discussion was divided into two subgroups: the first subgroup studied how different representation and modeling alternatives currently proposed can fit in a bigger picture of theory and technology interaction, while the second subgroup focused on contrasting current system implementations and the reasons behind such diverse class of available prototypes. We summarize the findings of these two groups and the future steps suggested by group members.

Joint work of: Das Sarma, Anish; de Keijzer, Ander; Deshpande, Amol; Haas, Peter J.; Ilyas, Ihab F.; Koch, Christoph; Neumann, Thomas; Olteanu, Dan; Theobald, Martin; Vassalos, Vasilis

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2009/1941

08421 Working Group: Imprecision, Diversity and Uncertainty: Disentangling Threads in Uncertainty Management

We report on the results of Workgroup 1 on "Imprecision, Diversity and Uncertainty". We set the scene by elaborating on where uncertainty comes from and what the ground truth is. In real world applications, the data observed may not be as expected: they may violate constraints, or, more generally, disagree with the anticipated model of the world. This leads to two orthogonal cases: The data may be erroneous, i.e. they must be corrected. Or, the model may outdated and must be adjusted to the data. After elaborating on this fundamental distinction, we address the issues of measuring uncertainty and exploiting uncertainty in real applications. We conclude with a list of challenges that should be addressed when dealing with uncertainty.

Keywords: Probabilistic databases, explanation component, transparenca, sources of uncertainty, presenting uncertainty

Joint work of: Spiliopoulou, Myra; van Keulen, Maurice; Lenz, Hans-Joachim; Wijsen, Jef; Renz, Matthias; Kruse, Rudolf; Stern, Mirco

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2009/1937

08421 Working Group: Explanation

This working group addressed the issue of explaining the results of an uncertainty information system to a user. For that, we structured the problem along three major queries: why, what, and how.

Keywords: Probabilistic databases, explanation component; transparenca, sources of uncertainty, presenting uncertainty

Joint work of: Aras, Hidir; Fuhr, Norbert; Hwang, Seung-won; de Keijzer, Ander; Klan, Friederike; Lenz, Hans-Joachim; Matthé, Tom; Schweppe, Heinz; Stern, Mirco; De Tré, Guy

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2009/1935

08421 Working Group: Lineage/Provenance

The following summary tries to capture a collection of state-of-the-art techniques and challenges for future work on lineage management in uncertain and probabilistic databases that we discussed in our working group. It was one half of a larger committee that we had initially formed, which then got split into two groups—one focusing on lineage as a means of explanation of data, and one focusing more on lineage usage in probabilistic databases (see also the "Explanation" working group report for more details on the first subgroup).

Keywords: Lineage and provenance, probabilistic databases, challenges

Joint work of: Das Sarma, Anish; Deshpande, Amol; Hubauer, Thomas; Ilyas, Ihab F.; König-Ries, Birgitta; Renz, Matthias; Theobald, Martin

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2009/1931

08421 Working Group: Report of the Probabilistic Databases Benchmarking

The results of the probabilistic database benchmark working group.

Keywords: Probabilistic databases, benchmark

Joint work of: Koch, Christoph; Haas, Peter J.; Lenz, H.-J.; Olteanu, Dan; Re, Christopher; van Keulen, Maurice; Pan, Jeff Z.

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2009/1936

08421 Working Group: Uncertainty and Trust

This report summarizes the findings of a working group on "Uncertainty and Trust" which met during Dagstuhl Seminar 08421 "Uncertainty Management in Information Systems". All participants of the working group are co-authors of this report. The aim of the working group was to analyse the relationship between trust and uncertainty in distributed reputation systems. We started by identifying sources and types of uncertainty in this context and investigated their relation to trust. After that we compiled a list of desirable properties of trust representations and finally determined open research challenges in the area.

Keywords: Trust, reputation, uncertainty, decentral system

Joint work of: Aras, Hidir; Beckstein, Clemens; Buchegger, Sonja; Dittrich, Peter; Hubauer, Thomas; Klan, Friederike; König-Ries, Birgitta; Wolfson, Ouri

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2009/1938

2 Plenary Talks

Trio: A System for Data, Uncertainty, and Lineage

Anish Das Sarma (Stanford University, US)

In the Trio project at Stanford, we are building a DBMS that supports data, its uncertainty, and its lineage as first-class interrelated concepts. In this talk, I will give an overview of Trio, describing the data model, query language, and the system. I shall highlight some past and ongoing research, and present possible directions for future work in uncertain data management.

Keywords: Uncertain data, lineage, Trio

Uncertain Data Management for Sensor Networks

Amol Deshpande (University of Maryland - College Park, US)

Much of the real-world data generated these days is inherently uncertain or probabilistic in nature. For instance, sensor data typically has some notion of quality attached to it – it could be the confidence of existence, trust, accuracy, a probability distribution over possible values, or a mix of these.

Similarly, when attempting to integrate heterogenous data sources (data integration) or extracting structured information from text (information extraction), the results are approximate and uncertain at best.

In this talk, I will present an overview of our approach to integrate statistical and probabilistic models into relational database systems so as to make it easy to manage and reason about uncertain data. I will first discuss our work on the MauveDB project that supports an abstraction called "model-based views" using which users can specify statistical models to be applied to streaming sensor data. The output of the modeling process is presented to the users as a relational table that can be queried using a declarative language. I will then present our ongoing work on query processing over uncertain relational databases, which occur naturally in many settings such as information extraction or may be the result of a probabilistic model-based view. We have developed a uniform, closed framework for representing and querying uncertain data based on concepts from probabilistic graphical models; I will present an overview of our framework, the challenges in query evaluation over uncertain data, and the algorithms we have developed for efficient query evaluation over large uncertain databases.

Vague Predicates, Probabilistic Rules and 4-Valued Logic for Probabilistic Databases

Norbert Fuhr (Universität Duisburg-Essen, DE)

This talk addresses three different issues that should be given more attention when developing new systems for uncertainty management, especially for the integration of information retrieval and database systems. First, vague (or fuzzy) predicates abstract from specific methods of text indexing, thus representing the appropriate logical level for the IR component, and also can be extended to arbitrary data types, especially multimedia. Second, probabilistic rules in probabilistic Datalog extend reasoning over uncertain facts by inferring uncertain statements from deterministic facts. Finally, 4-valued logic should be used when possibly contradicting facts from different sources have to be integrated.

A Monte Carlo Approach to Managing Uncertain Data

Peter Haas (IBM Almaden Center - San José, US)

There is an increasing need for tools that facilitate business decisionmaking in the face of uncertain data. The problem of data uncertainty is becoming acute, due to data integration, automated information extraction, data anonymization for privacy protection, and the growing importance of RFID and sensor data. Recently, in joint work between IBM Research and University of Florida, the MCDB relational database system has been developed for managing uncertain data, based on a Monte Carlo approach. This system can handle complicated real-world queries and data, and has an extensible and flexible uncertainty model, encapsulated via user-defined "value generation" (VG) functions. MCDB also allows sophisticated, data-intensive stochastic modeling and prediction to be performed close to the data. The key technical idea is to process a query plan once, but over "tuple bundles" that encapsulate possible worlds, rather than ordinary tuples; pseudorandom number seeds are used to compress such bundles whenever possible, in order to provide acceptable performance. We give an overview of the MCDB system, and also describe a recent effort to implement the MCDB functionality within the the setting of cloud computing, specifically, Hadoop, augmented with the Jaql query language. We focus on cloud computing not only because it is an increasingly popular and ubiquitous computing model, but also because it can potentially speed up the CPU-intensive MCDB computations via massive parallelism, and permits straightforward extension of MCDB functionality to certain types of non-relational data. Key challenges in this setting are to manage the pseudorandom number seeds that form the basis of the Monte Carlo computations, and to understand the tradeoffs between "inter-tuple" parallelism and "intra-tuple parallelism" (generating possible worlds for a single tuple in parallel). We also briefly describe our ongoing efforts to develop a realistic end-to-end business scenario, based on current work by IBM's AVATAR

project to develop principled algorithms for assigning probabilities to annotated data.

Keywords: Monte Carlo, probabilistic database, Hadoop, pseudorandom number generation

URank: Ranking Uncertain Data

Ihab Ilyas (University of Waterloo, CA)

Ranking and aggregation queries are widely exploited in data exploration, data analysis and decision making scenarios. While most of the currently proposed ranking and aggregation techniques focus on deterministic data, several emerging applications involve data that is unclean or uncertain. Ranking and aggregating uncertain (probabilistic) data raises new challenges with respect to query semantics and processing, which makes conventional methods inapplicable.

In this talk, I will introduce new formulations for ranking and aggregation queries in probabilistic databases. The new formulations are based on marriage of traditional ranking and aggregation algorithms with possible worlds semantics. In the light of these formulations, I will describe a generic processing framework supporting both query types, and leveraging existing query processing and indexing capabilities in current database systems. The framework encapsulates a state space model, and efficient search algorithms that compute query answers with optimality guarantees.

Keywords: Ranking, uncertain, probabilistic

MayBMS: A System for Managing Large Uncertain and Probabilistic Databases

Christoph Koch (Cornell University - Ithaca, US)

MayBMS is a state-of-the-art probabilistic database management system that has been built as an extension of Postgres, an open-source relational database management system. MayBMS follows a principled approach to leveraging the strengths of previous database research for achieving scalability. This talk describes the main goals of this project, the design of query and update language, efficient exact and approximate query processing, algorithmic and systems aspects, and applications, and gives an outlook at future work.

Keywords: Probabilistic databases, implementation, query languages, algorithms

See also: http://www.cs.cornell.edu/bigreddata/maybms/

An Overview of the Mystiq Probabilistic Database

Christopher Re (University of Washington, US)

In this talk, we describe our ongoing work at the University of Washington on the Mystiq system, a probabilistic database (pDB) that is motivated by applications in deduplication, information extraction and RFID. In particular, we discuss three techniques that form the technical core of query evaluation in Mystiq: (1) safe plans, a technique that allows us to evaluate SQL queries on probabilistic databases as fast as native SQL, (2) multisimulation, a technique that allows us to evaluate any SELECT-FROM-WHERE query efficiently and (3) materialized views for pDBs that allows GB scale SQL processing.

Keywords: Probabilistic relational data

Probabilistic Data Integration

Maurice van Keulen (University of Twente, NL)

In data integration efforts such as in portal development, much development time is devoted to entity resolution. Often advanced similarity measurement techniques are used to remove semantic duplicates or solve other semantic conflicts. It proofs impossible, however, to automatically get rid of all semantic problems. An often-used rule of thumb states that about 90% of the development effort is devoted to semi-automatically resolving the remaining 10% hard cases. In an attempt to significantly decrease human effort at data integration time, we have proposed an approach that strives for a 'good enough' initial integration which stores any remaining semantic uncertainty and conflicts in a probabilistic XML database. The remaining cases are to be resolved during use with user feedback.

We conducted extensive experiments on the effects and sensitivity of rule definition, threshold tuning, and user feedback on the integration quality. We claim that our approach indeed reduces development effort - and not merely shifts the effort - by showing that setting rough safe thresholds and defining only a few rules suffices to produce a 'good enough' integration that can be meaningfully used, and that user feedback is effective in gradually improving the integration quality.

Keywords: Data integration, entity resolution, probabilistic databases, data quality

Joint work of: Maurice van Keulen, Ander de Keijzer

3 Session Talks

On APIs for probabilistic databases

Lyublena Antova (Cornell University - Ithaca, US)

This talk targets challenges associated with managing updates and transactions on probabilistic databases. Traditional DBMSs provide APIs that provide access to the data on the logical level and hide physical representation details. Probabilistic database often employ compact ways to represent large sets of possible worlds, which adds additional layers of abstraction.

I will present a programming model for uncertain and probabilistic databases that is independent of representation details. Conceptually, we use the possible worlds semantics, and programs are independently evaluated in each world. We study a class of programs that appear to the user as if they are running in a single world rather than on a set of possible worlds.

We present an algorithm for efficiently verifying this property. We discuss how updates can be implemented in uncertain database management systems, and propose techniques for optimizing database programs.

Joint work of: Lyublena Antova, Christoph Koch

Contextual Information Processing in SmartWeb

Hidir Aras (Universität Bremen, DE)

Contextual Information is essential for adaptive intelligent dialogue systems. Understanding the context of the user can help to return better answers to natural language queries. Especially for mobile question answering (QA) systems quick and precise answers decide whether users will use such systems or not. In the SmartWeb project - whose goal was to realize a multi-modal mobile access to the Semantic Web Ű we identified a variety of contextual parameters that we integrated via ontological contextual models. The developed models were used to adapt the systems internal states to allow for context-aware behaviour. They also helped to reduce the needed dialogue turns in question answering. While ambiguous queries could not be answered satisfactory so far by existing systems, our models were able to cope with such pragmatic ambiguities and resolve them. The contextual models not only helped to enrich the semantic representations for incoming natural language queries but also disambiguated complex queries using a semantic coherence measure. De-contextualization was used to cope with underspecified natural language queries, context-specific recommendations or selection of appropriate semantic wrappers to extract relevant semantic structures from heterogeneous web sources.

Keywords: Contextual models, pragmatic knowledge, smartweb, question answering, web extraction

Joint work of: Hidir Aras, Robert Porzel and Rainer Malaka (University of Bremen); Berenike Loos, Vanessa Micelli and Hans-Peter Zorn (European Media Lab)

Full Paper:
http://smartweb.dfki.de/eng/papers_en.html#UB

Full Paper:
http://smartweb.dfki.de/eng/papers_en.html#eml

Using Reputation Systems to Reduce Behavior Uncertainty

Sonja Buchegger (Deutsche Telekom Laboratories, DE)

When making decisions about transactions with others, we need to deal with uncertainty and determine how much trust to place into a transaction to minimize the risk associated with a bad decision, i.e. the probability of occurance and the potential impact. Reputation systems aim at reducing the uncertainty over the behavior of others, be it people, services, products (as in recommendation systems) or nodes in self-organized networks, such as peer-to-peer or mobile ad-hoc networks. Reputation systems work on the premise of being able to predict future behavior by looking at past behavior. A good estimation of future behavior is not easy, the method of prediction needs to be adaptive and deal with erratic behavior, which is by definition not predictable. Yet, when there are no direct means of behavior enforcement, we have to try to get the best estimation we can to minimize expected risk. Our reputation system approach uses a combination of Bayesian estimation, integration of second-hand information, weighting, fading of past behavior, and immune-system inspired secondary response for repeated bad behavior. We find that reputation systems, when using information from different sources by leveraging on second-hand information can substantially improve the estimation of expected behavior based on the past. Reputation systems, however, introduce a different problem, that of dishonest reports, and thus have to find a balance between warding of liars and taking advantage of other participants' information.

Keywords: Reputation systems

Indexing Probability Density Functions in Parametric Form

Christian Böhm (LMU München, DE)

Our approach to similarity search and data mining on large databases of uncertain feature vectors is to represent each uncertain object by a multivariate probability density function (PDF).

In contrast to previous approaches, our idea is not to represent the PDFs discretely (e.g. by some histogram) but in their original, parametric form. Typical and widely used probability distributions such as Gaussian, Laplacian, Generalized Exponential or even Mixture Models can in this way be represented in a lossless way. Organizing objects in the parameter space (e.g. by mean and variance for Gaussian distributions) requires a careful reformulation of methods for data mining and indexing structures but is very space and time efficient. We will present in detail the Gauss-tree, an index structure for Gaussian PDFs but we will also introduce some advanced data analysis methods such as skylines or clustering methods on our notion of uncertain objects.

Expectation-Expectation: Second-order Expectation for Reducing Uncertainty in Contingent Systems

Peter Dittrich (Universität Jena, DE)

Exectations play an important role in social systems. Here I will focus on one specific type of expection called second-order expectation or expectation-expectation for short. In the context of social systems theory, it has been sugested that second-order expectations are cruicial for the formation of social order. I will discuss whether this mechanism might be also applied in technological systems consisting of many communicating agents and actors for reducing unvertainty. Simulation studies show that that the degree of order depends on how expectation-expectations are learned.

Keywords: Social order, systems theory, double condingency problem

Abductive Reasoning for Imperfect information

Thomas Hubauer (Siemens - München, DE)

This talk adresses the challenges of semi-automatic information-integration and abstraction in real-world domains. While there is no lack of approaches to this problem, there are often severe restrictions limiting the applicability in challenging domains, for example concerning limited expressivity (Bayes Nets) or problems in handling imperfect date (many logic-based approaches).

We propose the use of abductive reasoning to handle two prominent qualities of information imperfection, namely incompleteness (lack of information), and uncertainty (degree of belief in the truth oof some piece of information). Abductive inference schemes naturally handle missing data by allowing for assumptions, and weighting based on probabilities allows to integrate uncertain information both on instance and schema level.

I conclude by presenting ideas on additional information-teoretic and cost criteria for guiding the abductive process in the context of situation recognition, and for triggering information-gathering activities.

Keywords: Abductive reasoning, imperfect information, information integration

Analysis of conflicting information using argumentation

Anthony Hunter (University College London, GB)

Argumentation systems aim to reflect how conflicting information is used in cognition to construct and analyse arguments. So these systems involve identifying arguments and counterarguments relevant to an issue (e.g. What are the pros and cons for the safety of mobile phones for children?). They may also involve weighing, comparing, or evaluating arguments (e.g. What sense can we make of the arguments concerning mobile phones for children?) and they may involve drawing conclusions (e.g. A parent answering the question "Are mobile phones safe for my children?"). In addition, they may involve convincing an audience (e.g. A politician making the case that mobile phones should be banned for children because the risk of radiation damage is too great). In this talk, I cover some of the basics of the theory and implementation of logic-based argumentation systems and then cover some work we are doing on an application in analysing evidence from clinical drug trials on treatments for breast cancer.

Keywords: Argumentation, inconsistency

Fast Approximate String Searching in Databases

Ela Hunt (The University of Strathclyde - Glasgow, GB)

Approximate string searching uses database indexing only to a limited extent. I will present a new development in indexing for natural language, with application to web documents and tested in both client-server and P2P scenarios, and possibly extensible to biological string searching. I will first discuss the concept of the deletion neighbourhood and outline some complexity issues related to this idea. Then, I will move to the application areas where the concept proved to be beneficial. I will summarise the results of various performance tests with Wikipedia, Moby Dick, and natural language dictionaries, and then move on to a P2P DHT-based scenario which was the subject of another test. Finally, I will discuss possible extensions of this work, and the forthcoming tests with biological sequences.

Keywords: Database string indexing, approximate string search, deletion neighbourhood

Full Paper:
http://fastss.csg.uzh.ch/

Skyline ranking for uncertain data

Seung-won Hwang (POSTECH - Pohang, KR)

Skyline queries have been actively studied lately as they can effectively identify interesting candidate objects with low formulation overhead. In particular, this talk discusses supporting skyline ranking for the data with uncertainty, e.g., automatically extracted data with extraction uncertainty.

A Uncertainty Perspective on Qualitative Preference

Seung-won Hwang (POSTECH - Pohang, KR)

Collaborative filtering has been successfully applied for predicting a person's preference on an item, by aggregating community preference on the item. Typically, collaborative filtering systems are based on based on quantitative preference modeling, which requires users to express their preferences in absolute numerical ratings. However, quantitative user ratings are known to be biased and inconsistent and also significantly more burdensome to the user than the alternative qualitative preference modeling, requiring only to specify relative preferences between the item pair. More specifically, we identify three main components of collaborative filtering– preference representation, aggregation, and similarity computation, and view each component from a qualitative perspective. From this perspective, we build a framework, which collects only qualitative feedbacks from users. Our rating-oblivious framework was empirically validated to have comparable prediction accuracies to an (impractical) upper bound accuracy obtained by collaborative filtering system using ratings.

Keywords: Collaborative filtering, qualitative preference, uncertainty

Joint work of: Hwang, Seung-won; Lee, Mu-Woong

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2009/1932

Recommendation: A Less Explored Killer-App of Uncertainty?

Seung-won Hwang (POSTECH - Pohang, KR)

Due to the unprecedented amount of information available, it is becoming more and more important to provide personalized recommendations on data, based on past user feedbacks. However, available user feedbacks or ratings are extremely sparse, which motivates the needs for rating prediction. The most widely adopted solution has been collaborative filtering, which (1) identifies "neighboring" users with similar tastes and (2) aggregates their ratings to predict the ratings of

the given user. However, while each of such aggregation involves varying levels of uncertainty, e.g., depending on the distribution of ratings aggregated, which has not been systematically considered in recommendation, though recent study suggests such consideration can boost prediction accuracy. To consider uncertainty in rating prediction, this paper reformulates the collaborative filtering problem as aggregating community ratings into multiple predicted ratings with varying levels of certainty, based on which we identify top-k results with both high confidence and rating. We empirically study the efficiency and accuracy of our proposed framework, over a classical collaborative filtering system.

Keywords: Recommendation, uncertainty

Joint work of: Hwang, Seung-won; Roh, Jong-won

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2009/1933

Towards Quantifying Uncertainties of Selectivity Estimations

Andranik Khachatryan (Universität Karlsruhe, DE)

Selectivity estimation is crucial for Query Optimization. Misevaluating the selectivity of the query can result in a poor execution plan choice.

To aid the task of selectivity estimation, a concise summary of the data distribution is constructed (histograms and samples are the ones most commonly used).

There has been much work done to provide good estimations. However the estimations are never exact: there is always uncertainty involved. From the point of view of Query Optimizer, it is very important to know the degree of uncertainty associated with the estimate. In this presentation we quantify the uncertainty involved in selectivity estimations. Specifically, we show that probability distributions and prediction intervals of possible selectivities can be issued instead of a point estimates.

Keywords: Selectivity estimation, histograms, query optimization

Temporal Aspects in Exploratory Data Analysis

Rudolf Kruse (Universität Magdeburg, DE)

This talk targets challenges associated with temporal aspects in exploratory data analysis in automotive industry. New results concerning the topics mining temporal patterns, temporal dependencies in bayesian networks and visualisation of temporal dependencies via fuzzy rules are presented.

Keywords: Bayes networks, visual data mining

Spreadsheet Computation with imprecise and uncertain Data

Hans-Joachim Lenz (FU Berlin, DE)

We consider universal relations as flat tables having numeric data types which can be simply viewed as spreadsheets. For more details please refer to the extended abstract.

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2009/1939

Optimizing Imprecise Aggregation Queries in Large Distributed Networks

Thomas Neumann (MPI für Informatik - Saarbrücken, DE)

Information retrieval system have to deal with imprecise information routinely, for example due to uncertainty in data extraction processes or due to probabilistic relevance models. Queries are usually performed in a top-k style manner where the the k most 'relevant' results for a query are computed.

Unfortunately the standard techniques for top-k queries fail in a distributed or peer-to-peer setting, as network transfer is very costly and requires specialized algorithm.

In fact it seems to be infeasible to compute the exact answer in large networks: We could demonstrate that the uncertainty during query processing forces existing algorithms to send a very substantial fraction of the data over the network, which is prohibitive expensive.

We therefore propose two approaches: First, to use query optimization techniques to improve network traffic in general, and second, to use data-adaptive sampling to reduce the relevant part of the data. In our experiments this greatly reduced the query execution time while still offering high recall.

Using OBDDs for Efficient Query Evaluation on Probabilistic Databases

Dan Olteanu (University of Oxford, GB)

This talk addresses the problem of query evaluation for tuple independent probabilistic databases and conjunctive queries.

In the first part of the talk I show how this query evaluation problem can be approached as a construction problem for ordered binary decision diagrams (OBDDs): Given a query and a probabilistic database, we construct in polynomial time an OBDD such that the confidences of the answer tuples can be computed linearly in the size of that OBDD. This approach is applicable to a

large class of queries, including the hierarchical queries, i.e., the Boolean conjunctive queries without self-joins that admit PTIME evaluation on any tuple-independent probabilistic database, hierarchical queries extended with inequalities, and non-hierarchical queries on restricted databases.

In the second part of the talk I will present an efficient secondary-storage operator for exact confidence computation of hierarchical queries on tuple-independent probabilistic databases. This operator is based on the OBDD construction procedure discussed in the first part of the talk, and uses structural properties of the query and functional dependencies that hold on the database to decide on the number of scans of the answer tuples necessary to compute the confidences.

A case study on the TPC-H benchmark reveals that most TPC-H queries obtained by removing aggregations admit efficient evaluation using our operator. Experimental evaluation on probabilistic TPC-H data shows substantial efficiency improvements when compared to the state of the art. This operator is part of the engine of MayBMS, an open-source database management system for uncertain and probabilistic data.

Keywords: Probabilistic databases, query evaluation, secondary-storage algorithms, OBDDs, MayBMS

See also: Talk based on results published in Proc. of Int. Conf. on Scalable Uncertainty Management (SUM), 2008; Proc. of IEEE Int. Conf. on Data Engineering, 2009

Scalable Querying Services over Fuzzy Ontologies

Jeff Z. Pan (University of Aberdeen, GB)

Fuzzy ontologies are envisioned to be useful in the Seman- tic Web. Existing fuzzy ontology reasoners are not scalable enough to handle the scale of data that the Web provides.

In this paper, we propose a framework of fuzzy query lan- guages for fuzzy ontologies, and present query answering algorithms for these query languages over fuzzy DL-Lite on- tologies. Moreover, this paper reports on implementation of our approach in the fuzzy DL-Lite query engine in the ONTOSEARCH2 system and preliminary, but encouraging, benchmarking results. To the best of our knowledge, this is the first ever scalable query engine for fuzzy ontologies.

Keywords: Semantic Web, ontology, fuzzy, scalable, query

Full Paper:
http://www2008.org/papers/pdf/p575-panA.pdf

Uncertainty and Information Integration in Biomedical Applications

Claudia Plant (TU München - Klinikum Rechts der Isar, DE)

Due to the advances of imaging and high throughput technologies for data acquisition in biomedicine an increasing amount of data is produced. In many applications, data is associated with uncertainty, often due to specific properties and limitations of data aquisition, eg. the resolution of some imaging modality. In addition, integration of data from different sources is essential to optimally support the knowledge discovery process. This talk will introduce the challenges and chances of data integration in the context of two concrete applications in neurosciences and proteomics. Combining uncertain information from different sources may reduce uncertainty and thereby effectively support knowledge discovery.

Probabilistic Similarity Queries in Uncertain Databases

Matthias Renz (LMU München, DE)

Many modern applications have to cope with objects comprising vague and uncertain data. Example applications are location determination and proximity detection of moving objects, similarity search and pattern matching in sensor databases or personal identification and recognition systems based on video images or scanned image data. In the recent decade a lot of approaches that address the management and efficient query processing of uncertain data have been published. They mainly differ in the representation of the uncertain data, the distance measures, the types of queries, the query predicates and the representation of the result. This talk gives an overview of methods for effective and efficient similarity queries on uncertain data in feature databases. It especially emphasizes probabilistic similarity ranking methods that exploit the full information given by inexact object representations. Here, we assume the discrete uncertainty model where the uncertain point objects are represented by a set of alternative vectors which are assigned confidence values reflecting the likelihood that the point object is located at the corresponding vector position.

Keywords: Probabilistic query, similarity search, uncertain data, probabilistic ranking

Outlier detection and ranking based on subspace clustering

Thomas Seidl (RWTH Aachen, DE)

Detecting outliers is an important task for many applications including fraud detection or consistency validation in real world data.

Particularly in the presence of uncertain data or imprecise data, similar objects regularly deviate in their attribute values. The notion of outliers has thus to be defined carefully. When considering outlier detection as a task which is complementary to clustering, binary decisions whether an object is regarded to be an outlier or not seem to be near at hand. For high-dimensional data, however, objects may belong to different clusters in different subspaces. More fine-grained concepts to define outliers are therefore demanded. By our new OutRank approach, we address outlier detection in heterogeneous high dimensional data and propose a novel scoring function that provides a consistent model for ranking outliers in the presence of different attribute types. Preliminary experiments demonstrate the potential for successful detection and reasonable ranking of outliers in high dimensional data sets.

Keywords: Outlier detection, outlier ranking, subspace clustering, data mining

Joint work of: Thomas Seidl, Emmanuel Müller, Ira Assent, Uwe Steinhausen

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2009/1934

On the Expressiveness of Probabilistic XML Models

Pierre Senellart (Telecom Paris Tech, FR)

Various known models of probabilistic XML can be represented as instantiations of the abstract model of p-documents. In addition to ordinary nodes, p-documents have distributional nodes that specify the possible instances and their probabilistic distribution.

One can specify particular families of p-documents by allowing specific kinds of distributional nodes as well as structural constraints on the placement of such nodes in a p-document; some of the resulting families provide natural extensions and combinations of previously introduced probabilistic XML models.

The focus of this talk is on the expressive power of families of p-documents. In particular, two main issues are studied. The first is the ability to (efficiently) translate a given p-document of one family into another family. The second is closure under updates, namely, the ability to (efficiently) represent the result of updating the instances of a p-document of a given family as another p-document of that family. For both issues, we distinguish two variants corresponding to value-based and object-based semantics of p-documents.

Keywords: XML, probabilistic databases, probabilistic XML, expressiveness, updates

Joint work of: Serge Abiteboul, Benny Kimelfeld, Yehoshua Sagiv, Pierre Senellart

Probabilistic databases based on probabilistic graphical models

Pritviraj Sen (University of Maryland - College Park, US)

Probabilistic graphical models are a popular and well studied framework for compact representation of joint probability distributions involving many interdependent variables, and for efficient reasoning about such distributions. In this talk, we focus on the connections between query evaluation over probabilistic databases, probabilistic inference and large-scale structured probabilistic graphical models (such as probabilistic relational models). In particular, we will discuss how to 1) represent uncertainty in databases via the use of random variables and probability distributions, 2) how to cast a query evaluation problem for probabilistic databases as a marginal probability computation problem and 3) explore the connections between query evaluation in probabilistic databases and lifted inference (a recent area of research in machine learning). Our hope is that by stimulating discussion between the two fields of databases and machine learning we will promote synergy which in turn, will lead to new discoveries that will be beneficial to researchers from either field.

Keywords: Probabilistic databases, probabilistic graphical models, probabilistic inference, lifted inference

Joint work of: Prithviraj Sen, Amol Deshpande, Lise Getoor

Dealing with uncertainty in models over data

Myra Spiliopoulou (Universität Magdeburg, DE)

As data proliferate and information overload occurs, people rely increasingly on *derived* information, i.e. on models and patterns extracted from the data, rather than on the data themselves. However, to what extend can we rely on a derived model? There are many factors that can challenge the validity of a model over the data and thus question conclusions and decisions taken upon the model. Among them:

- The data used to derive the model are no more representative of the population because of endogenous factors: For example, customer segments derived 20 years ago for customers in Europe should be reconsidered because the population of Europe is becoming older.
- The population originally represented in the model has changed because of the model extraction process itself: For example, a marketing campaign has been designed for specific customer profiles, but this campaign has lead to cannibalization effects among existing products.

So, how certain are the conclusions we draw from a model? And are there any indicators or measures about the certainty, resp. uncertainty, of such conclusions?

Subject of this presentation is capturing change in derived models, with emphasis on how model change over time, i.e. their validity becomes uncertain as time passes. Data are observed as streams and study how a model fits an evolving stream. For this task, a distinction must be made between models in supervised and unsupervised learning.

- In supervised learning, there is a ground truth, so we can conclude that the model does not fit the data any more, i.e. there is a "concept drift". The task is then to detect the drift and adapt the model.
- In unsupervised learning, there is no ground truth. One can still conclude that the model does not fit the data, but the process of deriving the model is itself uncertain. Hence, the task is not only to adapt the model but also to capture and interpret changes upon the model. We discuss how this task is dealt with in stream clustering.

There are many open questions on how uncertain data and uncertain models should be dealt with. Some questions for this seminar:

1. We derive models from data and we know that their lifetime (in terms of validity) is limited. Which properties of the data can we exploit or monitor to conclude that a model is becoming invalid, or, more precisely it is changing – and how it is changing?
2. Uncertainty on the data may refer to systematic and non-systematic perturbations. Non-systematic perturbations are of less interest for model extraction; they are noise. However, how sensitive is a model to noise?
   And how can we distinguish between data that are uncertain because of noise and those that exhibit concept drift?

Keywords: Model evolution, stream mining, temporal mining, model comparison

TrioOne: Layering Uncertainty and Lineage on a Conventional DBMS

Martin Theobald (Stanford University, US)

In the Trio project at Stanford, we are developing a new kind of database management system - one that handles data, uncertainty of the data, and data lineage together in a fully integrated manner. Some of the application domains targeted by Trio are data cleaning and integration, information extraction, and scientific data management.

Our first system prototype, dubbed Trio-One, is primarily layered on top of a conventional relational DBMS. From the user and application standpoint

Trio-One appears to be a "native" implementation of the Trio data model, query language, and other features. However, Trio-One encodes the uncertainty and lineage present in Trio's data model in conventional relational tables, and it uses a rewrite-based approach for most data management and query processing. A small number of stored procedures are used for specific functionality and increased efficiency.

The core system is implemented in Python and mediates between the underlying relational DBMS (currently the PostgreSQL open-source DBMS) and Trio interfaces and applications. The Python layer presents a simple Trio API that extends the standard Python DB 2.0 API for database access (Python's analog of JDBC). The Trio API accepts TriQL queries in addition to regular SQL, exposes lineage tracing, on-demand confidence computations, as well as some other Trio-specific features. Using the Trio API, we built a generic command-line interactive client similar to that provided by most DBMS's, and a full-featured graphical user interface called TrioExplorer.

Consistent Query Answering under Primary Key Violations

Jef Wijsen (Université de Mons, BE)

Database integrity constraints express properties that, ideally, should be satisfied at all times.

Nevertheless, data inconsistency is a phenomenon that often occurs in practice, especially in the context of data integration. The construct of database repair provides an elegant way to cope with database inconsistency. Intuitively, a repair of an inconsistent database db is a consistent database that differs from db in some minimal way. In general, an inconsistent database can have multiple repairs, each of which is equally possible. If there is more than one repair, then the set of all repairs represents incomplete (or uncertain) information.

The aim of consistent query answering is to get correct information out of an inconsistent database. Given an inconsistent database db, the goal is to determine information that is true irrespective of the way of repairing. More precisely, a Boolean query is said to be consistently true (or certain) if it evaluates to true on every repair of db. Consistent query answering has been the subject of much research since 1999.

The first part of this talk will discuss several aspects of database repairing and consistent query answering in a more general way. The second part will focus on practical approaches for consistent answering to conjunctive queries under primary key violations.

Keywords: Database repairing, consistent query answering

Uncertainty in Moving Object Databases

Ouri Wolfson (Univ. of Illinois - Chicago, US)

This work addresses the problem of managing Moving Objects Databases (MOD) which capture the inherent imprecision of the information about the moving object's location at a given time. We deal systematically with the issues of constructing and representing the *trajectories* of moving objects and querying the MOD.

In the first part of the talk we address the problem of determining when the location of a moving object in the database should be updated. We answer this question by proposing an information cost model that captures uncertainty, deviation, and communication cost, and then solving an optimization problem.

In the second part of the talk we propose to model an uncertain trajectory as a 3D cylindrical body and we introduce a set of novel but natural spatio-temporal *operators* which capture the *uncertainty* and are used to express spatio-temporal range queries. We devise and analyze algorithms for processing the operators and demonstrate that the model incorporates the uncertainty in a manner which enables efficient querying, thus striking a balance between the modeling power and computational efficiency.

See also: G. Trajcevski, O. Wolfson, K. Hinrichs, S. Chamberlain, "Managing Uncertainty in Moving Objects Databases", ACM Transactions on Database Systems (TODS), 29(3), Sept. 2004, pp. 463-507.

4 Demos

Fast Approximate String Searching in Databases

Ela Hunt (The University of Strathclyde - Glasgow, GB)

Approximate string searching uses database indexing only to a limited extent. I will present a new development in indexing for natural language, with application to web documents and tested in both client-server and P2P scenarios, and possibly extensible to biological string searching. I will first discuss the concept of the deletion neighbourhood and outline some complexity issues related to this idea. Then, I will move to the application areas where the concept proved to be beneficial. I will summarise the results of various performance tests with Wikipedia, Moby Dick, and natural language dictionaries, and then move on to a P2P DHT-based scenario which was the subject of another test. Finally, I will discuss possible extensions of this work, and the forthcoming tests with biological sequences.

Keywords: Database string indexing, approximate string search, deletion neighbourhood

Full Paper:
http://fastss.csg.uzh.ch/

MayBMS in Action

Dan Olteanu (University of Oxford, GB), Christoph Koch (Cornell University)

We will demonstrate MayBMS, a probabilistic database management system built as an extension of PostgreSQL. To date, MayBMS supports a powerful query language for processing and transforming uncertain data, space-efficient representation and storage, efficient query evaluation based on mature relational technology, and updates. MayBMS is publicly available at www.sourceforge.net under GPL license.

Joint work of: MayBMS Team

TrioOne: Layering Uncertainty and Lineage on a Conventional DBMS

Amish Das Sarma (Stanford University), Martin Theobald (Stanford University)

In the Trio project at Stanford, we are developing a new kind of database management system - one that handles data, uncertainty of the data, and data lineage together in a fully integrated manner.

Some of the application domains targeted by Trio are data cleaning and integration, information extraction, and scientific data management.

Our first system prototype, dubbed Trio-One, is primarily layered on top of a conventional relational DBMS. From the user and application standpoint Trio-One appears to be a "native" implementation of the Trio data model, query language, and other features. However, Trio-One encodes the uncertainty and lineage present in Trio's data model in conventional relational tables, and it uses a rewrite-based approach for most data management and query processing. A small number of stored procedures are used for specific functionality and increased efficiency.

The core system is implemented in Python and mediates between the underlying relational DBMS (currently the PostgreSQL open-source DBMS) and Trio interfaces and applications. The Python layer presents a simple Trio API that extends the standard Python DB 2.0 API for database access (Python's analog of JDBC). The Trio API accepts TriQL queries in addition to regular SQL, exposes lineage tracing, on-demand confidence computations, as well as some other Trio-specific features. Using the Trio API, we built a generic command-line interactive client similar to that provided by most DBMS's, and a full-featured graphical user interface called TrioExplorer.

Management of Imprecise, Incomplete and Uncertain Metric Data

Hans-Joachim Lenz (FU Berlin)

The demo of user-friendly software showed three approaches how to resolve the problem of "random" metric data given a system of linear and non-linear balance equations which may have missing (or null) values, outliers and measurement errors. QUANTOR - Schmid (1976) - uses a generalized least squares approach under the hypothesis of Gaussian distributed error variables. It approximates non-linear relationships by a first order Taylor approximation. Relaxing this assumptions and allowing for cross-correlation and any kind of finite parametric probability distributions leads to a MCMC approach implemented as MoSim by Köppen (2008). Finally, relaxing density functions and substituting them by (mostly triangle) membership functions leads to Fuzzy Logic - Lotfy Zadeh (1965). The implementation embedded into FuzzyCalc is due to Lenz and Müller (2003).

IMPrECISE: Good-is-good-enough Data Integration

Maurice van Keulen (University of Twente, NL)

The IMPrECISE system is a probabilistic XML database system which supports near-automatic integration of XML documents. What is required of the user is

to configure the system with a few simple knowledge rules allowing the system to sufficiently eliminate nonsense possibilities. We demonstrate the integration process under conditions with varying degrees of confusion and different sets of rules.

Even when an integrated document still contains much uncertainty, it can be queried effectively. The system produces a sequence of possible result elements ranked by likelihood. User feedback on query results further reduces uncertainty which in a sense continues the semantic integration process incrementally. We demonstrate querying on integrated documents and measure answer quality with adapted precision and recall measures. The user feedback mechanism has not been implemented, hence cannot be demonstrated yet.

IMPrECISE has been implemented as an XQuery module for the XML DBMS MonetDB/XQuery. Therefore, the demo also illustrates the power of this XML DBMS and of XQuery as both a query and programming language.

Joint work of: Maurice van Keulen, Ander de Keijzer

Full Paper:
http://eprints.eemcs.utwente.nl/11232/

See also: de Keijzer, A. and van Keulen, M. (2008) IMPrECISE: Good-is-good-enough data integration. In: Proceedings of the 24th International Conference on Data Engineering (ICDE2008), 7-12 April 2008, Cancun, Mexico. pp. 1548-1551. IEEE Computer Society Press. ISBN 978-1-4244-1837-4